图书情报工作
第 65 卷 第 3 期　2021 年 2 月

“一带一路”多语种共享型数据库的跨语言检索功能分析与开发策略*

司莉　周璟
武汉大学信息管理学院　武汉 430072

摘　要：[目的/意义]要实现“一带一路”多语种共享型数据库资源的有效利用，必须解决跨语言检索问题，基于已建“一带一路”数据库检索功能调查结果，分析“一带一路”多语种共享型数据库检索功能需求，以调研跨语言检索平台为视角，为“一带一路”多语种共享型数据库的跨语言检索功能设计与开发提供参考。[方法/过程]采用文献调研法和网络调研法，选取 11 个国内外典型的跨语言检索平台，从跨语言检索方法、跨语言翻译实现方法、检索功能设置、检索结果呈现、界面与检索支持语种 6 个方面进行分析，总结其实现方法。[结果/结论]为“一带一路”多语种共享型数据库的跨语言检索功能设计与开发提出策略：应采用基于神经网络机器翻译的提问式 – 文献翻译方法，实现多种检索功能，应用可视化技术呈现检索结果，提供多语言检索界面和资源。

关键词：“一带一路”数据库　多语种　跨语言检索
分类号：G250.74
DOI：10.13266/j.issn.0252 – 3116.2021.03.003

1　引言

自 2013 年 9 月习总书记提出“一带一路”倡议以来，世界各国和国际组织积极响应，截至 2020 年 1 月底，已有 138 个国家和 30 个国际组织与中国签署了 200 份共建“一带一路”合作文件[1]。多语种共享型数据库作为多元文化交流和信息资源共享的基础设施平台，能为建设“一带一路”多语言信息资源保障体系提供有效支撑。近年来，国内政府部门、高校、科研机构和企业已建立了多个“一带一路”数据库，经笔者调研，仅有 4 个平台提供多语言信息服务。跨语言检索指可用一种语言进行提问，检索出另一种或多种语言信息的信息检索技术[2]，基于跨语言检索的多语言信息服务帮助用户使用自己熟悉的语言文字，了解、浏览或者阅读其他语种信息资源的内容，能扩大不同语种信息资源共享范围[3]。当前“一带一路”数据库尚未真正实现跨语言检索，无法满足不同母语背景的用户对多语言信息资源的检索需求。在多语言信息服务方面，如何实现“一带一路”多语种共享型数据库的跨语言检索，服务“一带一路”沿线国家多语言信息需求，是当前亟需解决的重要课题。因此，本文在了解已建“一带一路”数据库检索功能的基础上，分析“一带一路”多语种共享型数据库检索功能需求，并对国内外典型的跨语言检索平台进行调查，分析其功能设计与开发实践，从而提出“一带一路”多语种共享型数据库的跨语言检索功能开发策略。

2　相关研究

如何实现“一带一路”多语种共享型数据库的跨语言检索功能，解决该数据库的多语种问题是本文的研究目标。因此，从跨语言检索翻译方法及其实现与“一带一路”数据库的多语种问题两个方面对相关研究进行梳理。

2.1　跨语言检索翻译方法及其实现

2.1.1　跨语言检索翻译方法

跨语言检索翻译方法有提问式翻译方法、文献翻

* 本文系教育部哲学社会科学研究重大课题攻关项目“‘一带一路’沿线国家多语种、共享型经济管理数据库建设研究”（项目编号：19JZD020）研究成果之一。
作者简介：司莉（ORCID：0000 – 0003 – 1028 – 8338），教授，博士，博士生导师，E-mail：lsiwhu@163.com；周璟（ORCID：0000 – 0002 – 6918 – 7163），硕士研究生。
收稿日期：2020 – 06 – 26　修回日期：2020 – 11 – 28　本文起止页码：20 – 27　本文责任编辑：王传清

译方法、提问式-文献翻译方法、中间语言翻译方法、非翻译方法,其中提问式翻译方法、文献翻译方法、提问式-文献翻译方法是目前主流翻译方法[4-5]。

提问式翻译方法是将提问式语种转换成文献目标语种,然后再进行单语言检索;文献翻译方法是将源文献语种转换成提问式语种,即不对提问式进行翻译,而是将集合中的文献翻译成与提问式语种一致的语言[6]。提问式-文献翻译方法综合两者优点,减少用户翻译成本的同时提高检索服务的质量,是目前实现跨语言检索比较理想的选择[7]。

2.1.2 跨语言翻译实现方法

跨语言翻译的实现方法有基于机器翻译的方法、基于语料库的方法、基于字典/词典的方法、字典/词典与语料库混合方法、基于本体的方法[6,8]。

基于机器翻译的方法是使用机器翻译系统进行翻译的方法。P. Iswarya 和 V. Radha 使用基于规则和统计的混合机器翻译系统,开发了跨语言文本检索系统,提高了翻译准确率和工作效率[9]。基于语料库的方法是将同一信息或同一主题的信息用两种或多种语言进行描述,并由人工或计算机建立不同语种间信息联系集合的方法[10]。R. Rahimi 等使用基于概率分布的模型提取源语言与目标词的对应关系,在可比语料库中构建翻译模型,为低频单词提供了更可靠的翻译[11]。基于字典/词典的方法利用机读字典/词典,将用户提交的检索式翻译成目标语种进行检索[12]。O. F. W. Onifade 等提出了一种具有双重概念驱动的文档聚类技术的模糊双语词典,来扩展词典翻译模型[13]。字典/词典与语料库混合方法结合上述两者优点,首先使用字典/词典对提问式进行翻译,然后使用专业语料库净化模糊不清的结果。J. Vilares 等基于提问式翻译方法,使用并行语料库自动生成的双语机器可读 N-gram 字典进行翻译,然后执行单语文本检索[14]。基于本体的方法指在语义层面翻译提问式,对检索对象进行语义处理,分析该语义段落中的潜在目标对象和查询请求的语义相关性,最后进行匹配[15]。孙玥莹等提出了一种基于领域知识库的科技术语信息匹配模型,并结合语言学特征、领域信息以及长短时记忆网络语言模型来挑选最合适译文[16]。

2.2 "一带一路"数据库的多语种问题

"一带一路"数据库的多语种问题给信息资源建设和数据库检索服务提出了新的挑战。于施洋等认为在数据采集和处理过程中,多语种问题是"一带一路"数据归集与普通数据库建设的最大差别之处[17]。已建"一带一路"数据库普遍缺乏小语种信息资源,尚未实现多语种资源的组织和整合[18],无法为跨语言检索功能提供资源保障,进而影响一带一路沿线国家间的信息资源共享。严丹等提出应关注和引进多语种资源,特别是中小国家语种原版资料,构建多语种、跨学科、多来源的"一带一路"信息资源体系[19-20]。梁昊光等认为加快建设基于多语种识别、多语言感知等语言技术的"一带一路"多语言云服务平台,提供基础数据资源和技术支撑[21],是"一带一路"数据库建设和开发的重要环节。但从理论和实践的角度来说,跨语言检索功能分析和开发仍是"一带一路"数据库建设的空白环节。

3 "一带一路"多语种共享型数据库检索功能需求分析

3.1 资源特点与用户需求

在资源特点方面,"一带一路"多语种共享型数据库涉及多语种、多类型、多领域、多来源的信息资源。信息资源语种多样,且部分语种较为小众,普及率较低,如柬埔寨语、匈牙利语、老挝语、波兰语、塞尔维亚语、越南语等;信息资源类型涵盖政策法规、统计数据、指数数据、研究与科技报告、新闻资讯、期刊与报纸、学位论文、著作、年鉴、经济管理重要工具书、专利文献、标准等;"一带一路"专题信息资源涉及政治、经济、文化、法律、国家安全等多个研究领域;信息资源来源包括各国政府机构、国际组织、科研院校、企业、数据库、权威媒体、权威智库和互联网等。

在用户需求方面,"一带一路"相关研究的信息资源需求主要集中于"一带一路"沿线国家和地区宏观信息需求、各国媒体发布的新闻报道和舆情信息需求、各国专业性领域的多语种信息需求以及对多语种、跨学科的学术资源和科研信息的需求[19]。可见,"一带一路"多语种共享型数据库需重点针对文献语种、文献类型、研究领域和文献来源等不同维度整合原始信息资源。"一带一路"沿线涉及 138 个国家和地区,涵盖百余种语种,数据库用户目前只能依赖自身多语言信息识读和理解能力,或借助外部翻译工具来获取小语种的原始信息资源。为帮助"一带一路"多语种共享型数据库用户理解与获取多语种信息资源,应配置符合用户检索需求的跨语言检索功能。

3.2 数据库检索功能调查

为了解"一带一路"多语种共享型数据库用户的

检索功能需求，对已有"一带一路"数据库检索功能进行调查，调查对象具体见专题文章《"一带一路"专题数据库建设调查与发展分析》表1部分，调查结果如下：

3.2.1 尚未实现跨语言信息检索

目前，尚未发现有"一带一路"数据库提供基于跨语言检索的多语言信息服务，已建"一带一路"数据库中，仅有丝路科技知识服务系统能够实现部分元数据层面的跨语言检索，其文献资源的题名、关键词和摘要通常包含英语或中英双语的翻译版本，如"题名"和"Alternate Title""英语摘要"和"Abstract from Author"分别用英语和源语言说明。当用户提交不同源语言的检索式时，系统通过提问式翻译方法，将检索式机器翻译结果与资源元数据进行匹配，但无法满足不同母语背景的用户对多语言信息资源全文的检索需求。信息检索中的语言障碍导致"一带一路"多语种资源难以被发现和利用。

3.2.2 大多仅支持简单检索功能

大多"一带一路"数据库仅支持简单检索，仅有27%的数据库支持高级检索，目前没有"一带一路"数据库支持专家检索，可用的运算符和可检字段较少，无法满足专业领域科研人员的检索需求。针对检索结果限定功能，52%的"一带一路"数据库支持从文献类型、文献主题、国别、发表年份等维度对检索结果进行精确，12%的数据库支持二次检索功能。可见，"一带一路"数据库在专家检索功能、检索结果限定功能、二次检索功能上仍有待改善。

3.2.3 检索结果呈现形式单一

检索结果的排序和可视化有利于用户快速掌握检出资源的概况和特征，准确定位自己所需的资源。仅有24%的"一带一路"数据库支持检索结果排序，这些数据库均可根据发表时间进行排序，此外，"中国一带一路网""列国志数据库""一带一路资源中心数据库"还可根据相关性进行排序。在检索结果可视化方面，"一带一路统计数据库"具备大量业务数据和统计数据，用户可将检索结果可视化，定制统计图，而其他数据库无法对检索结果进行可视化。目前"一带一路"数据库的检索结果呈现形式较为单一，会降低检索效率和用户体验。

3.2.4 大多不支持多语言界面

"一带一路"数据库的用户母语背景多样，超过单一语言或主流语言社群信息服务的范畴，而目前大多"一带一路"数据库不支持多语言界面，使得不能识读和理解其他语种的用户易产生认知障碍。其中，60%的"一带一路"数据库仅支持中文；9%的数据库仅支持英语；19%的数据库同时支持中文与英语；除此之外，有4个数据库支持包括中英在内的3种及以上语种，占比12%，其中美国EBSCO公司的数据库支持英语、日语、韩语、德语等30种语言界面，"中国一带一路网"支持中文、英语、俄语、法语、西班牙语、阿拉伯语6种联合国官方语言界面，"丝路科技知识服务系统"支持中文、英语、俄语、西班牙语4种语言界面，"新华丝路网"支持中文、英语、意大利语3种语言界面。大多数据库涵盖语种偏少，缺乏多语言界面，不利于用户以熟悉的语言文字为工具，了解、浏览或者阅读其他语种信息资源的内容，阻碍了不同语种信息资源的传播与利用[3]。

基于此，"一带一路"多语种共享型数据库需实现跨语言信息检索，设置简单检索、高级检索和专家检索功能，支持检索结果排序、分类限定和可视化，支持多语言界面等。其中，基于跨语言检索的多语言信息服务是"一带一路"数据库信息服务的关键环节，也是"一带一路"数据库检索功能建设的难点，目前已建"一带一路"数据库尚未有可借鉴的成熟经验。因此，需要吸取国内外跨语言检索平台的建设经验，建设和完善"一带一路"多语种共享型数据库的跨语言检索功能。

4 跨语言检索平台的调查分析

4.1 调查对象的选取

目前已有的国内外跨语言检索平台建设完善，可为"一带一路"数据库的跨语言检索功能开发提供参考。笔者参考李月婷和司莉提出的多语言信息组织模式[22]，使用网络调研法和文献调研法选取了11个跨语言检索平台作为调查对象，包括3个跨语言数据库、3个学科信息门户、2个搜索引擎和3个数字图书馆项目，具体如下：

（1）跨语言数据库。OECD iLibrary是以经济合作发展组织提供的信息资源为基础建立的数据库。IMF eLibrary是经济数据和分析报告数据库。AIpatent是南京深思得信息科技有限责任公司开发的专利情报检索系统。

（2）学科信息门户。WorldWideScience是一个跨语言、跨库科技文献检索平台，其资源涵盖70多个国家和地区，约100个数据库和门户网站，5亿多个网页的科学信息。丝路科技知识服务系统由西安交通大学

国际工程科技知识中心研发,是面向"一带一路"沿线需求的工程科技知识服务平台。石油石化大数据知识服务平台是基于我国石油石化行业建设的个性化知识服务体系。

(3)搜索引擎。2lingual Google Search 允许用户使用两种语言进行 Google 搜索,即输入检索式后,可选择任意两种语言获取搜索结果。搜狗海外搜索应用了搜狗机器翻译系统,可为用户提供英语原文、中文译文、中英双语 3 个版本的搜索结果。

(4)数字图书馆项目。世界数字图书馆(World Digital Library,WDL)是由美国国会图书馆、联合国教科文组织等发起的人类历史文化遗产数字图书馆项目。国际儿童数字图书馆(International Children's Digital Library,ICDL)由美国国家科学基金赞助,是马里兰大学和互联网档案馆合作研发的儿童数字图书馆项目,收录了反映不同时期、地域、文化和语言版本的数字化文学作品。欧洲数字图书馆(Europeana)是受欧盟委员会委托,由欧洲基金会主办的欧洲数字文化遗产项目,该项目涵盖了 1 500 多个博物馆、档案馆和图书馆的馆藏资源,提供 5 300 万数字对象访问权限。

4.2 跨语言检索平台调查结果

跨语言检索平台所要实现的核心步骤是翻译和检索。笔者从跨语言检索方法、跨语言翻译实现方法、检索功能设置、检索结果呈现、界面支持语种、检索支持语种 6 个维度对各跨语言检索平台进行调查,调查结果如表 1 所示:

表 1 国内外跨语言检索平台调查结果

平台类型	平台名称	跨语言检索方法	跨语言翻译实现方法	检索功能设置	检索结果呈现	界面支持语种(数量)	检索支持语种(数量)
跨语言数据库	OECD iLibrary[23]	文献翻译方法	机器翻译	简单检索、高级检索	检索结果排序、调整检索范围、资源格式选择、引文导出、分享至社交平台与邮件、保存检索式、查看检索历史、获取权限显示	英语、法语、日语(3 种)	同左
	IMF eLibrary[24]	文献翻译方法	机器翻译	简单检索、高级检索	检索结果排序、调整检索范围、二次检索、分享至社交平台与邮件	英语、西班牙语(2 种)	英语、法语、西班牙语(3 种)
	AIpatent[25]	提问式翻译方法	神经网络机器翻译	简单检索、高级检索、概念检索、自助检索	调整检索范围、二次检索	中文、英语、日语(3 种)	同左
学科信息门户	WorldWide Science[26]	提问式翻译方法	机器翻译	简单检索、高级检索	检索结果排序、调整检索范围、可视化、创建跟踪、分享至邮件、加入"我的图书馆"	英语(1 种)	中文、英语、俄语、法语、西班牙语、阿拉伯语、德语、日语、韩语、葡萄牙语(10 种)
	丝路科技知识服务系统[27]	文献翻译方法	机器翻译	简单检索、高级检索	检索结果排序、调整检索范围、引文导出、加入收藏	中文、英语、俄语、阿拉伯语(4 种)	中文、英语、俄语、法语、西班牙语、阿拉伯语(6 种)
	石油石化大数据知识服务平台[28]	文献翻译方法	机器翻译、人工翻译	简单检索、高级检索、分类检索、专家检索	检索结果排序、调整检索范围、二次检索、引文导出、可视化、查看检索历史	中文(1 种)	中文、英语
搜索引擎	2lingual Google Search[29]	提问式翻译方法	机器翻译	简单检索	双语检索结果分列显示	英语(1 种)	中文、英语、俄语、法语、西班牙语、阿拉伯语、保加利亚语、加泰罗尼亚语等(37 种)
	搜狗海外搜索[30]	提问式翻译方法	神经网络机器翻译	简单检索	选择显示原文、译文、双语、相关检索推荐	中文(1 种)	中文、英语(2 种)
数字图书馆	WDL[31]	文献翻译方法	机器翻译	简单检索	调整检索范围、选择结果显示方式、分享至社交平台与邮件	中文、英语、俄语、法语、西班牙语、阿拉伯语、葡萄牙语(7 种)	同左
	ICDL[32]	文献翻译方法	机器翻译、人工翻译	简单检索、高级检索	调整检索范围	英语、俄语、法语、西班牙语、蒙古语(5 种)	中文、英语、俄语、法语、西班牙语、阿拉伯语、波斯语等(18 种)
	Europeana[33]	文献翻译方法	机器翻译、基于语境词表	简单检索	调整检索范围、选择结果显示方式、分享至社交平台与邮件	英语、俄语、法语、西班牙语、保加利亚语、加泰罗尼亚语、捷克语等(26 种)	同左

4.2.1 跨语言检索方法

所调查的平台中,7 个平台采用文献翻译方法,占比为 64%,分别是 OECD iLibrary、IMF eLibrary、丝路科技知识服务系统、石油石化大数据知识服务平台、WDL、ICDL、Europeana。其中,IMF eLibrary、WDL、Europeana、ICDL 采用翻译待检索文献元数据的方法,方便用户了解每条资源的基本信息,并可采用对应语言检索出相关资源,尤其是 WDL、Europeana、ICDL 数字图书馆多为非文本馆藏,仅需提供馆藏元数据描述及其翻译;OECD iLibrary 数据库资源多为统计数据和分析报告,翻译工作量小,发布语言版本多,并直接提供部分文献全文的多语言翻译版本。搜索引擎待检索资源主要是网络资源,资源数量多且类型丰富,所以选择了目前最经济、工作量最小的提问式翻译方法,有 4 个平台采用提问式翻译方法,占比为 36%,分别是 AIpatent、WorldWideScience、2lingual Google Search、搜狗海外搜索。

4.2.2 跨语言翻译实现方法

目前跨语言翻译实现方法主要是机器翻译。所调查的平台均采用机器翻译方法,机器翻译速度远胜于人工翻译,但错误率仍较高。为提升机器翻译在具体应用场景的准确率,特别是商业合同、法律条文、专利文献等资源的翻译,AIpatent 和搜狗海外搜索采用了神经网络机器翻译技术,石油石化大数据知识服务平台和 ICDL 采用人工辅助翻译的方法,Europeana 采用了基于语境词表的翻译方法。其中 ICDL 面向世界各国儿童用户,对翻译的准确性、流畅性和趣味性要求较高,且资源主要为面向儿童的绘本类书籍,翻译工作量小,因此设置了专门的翻译志愿小组进行人工辅助翻译,负责翻译网站界面、基本书目信息、摘要和整本书籍,并由审核志愿小组进行校对。

4.2.3 检索功能的设置

所调查的平台均设置简单检索功能,有 7 个平台提供高级检索功能,占比为 64%,分别是 OECD iLibrary、IMF eLibrary、AIpatent、WorldWideScience、丝路科技知识服务系统、石油石化大数据知识服务平台、ICDL,且提供了检索功能使用指南,方便用户进行限定字段检索,或使用逻辑算符、位置算符和截词符进行组配检索。AIpatent 还提供概念检索功能,方便用户自主检索专利信息,用户可在概念检索编辑框输入发明专利技术交底书或专利全文,系统进行机器翻译和关键词提取;用户还可对关键词进行调整,进行二次检索,快速检索到符合用户需求的专利信息。

4.2.4 检索结果的呈现

对检索结果进行排序和范围调整是检索平台的基本功能。所调查的平台中,有 9 个平台支持通过“语言”“国别”“资源类型”“著者”等维度调整检索范围,占比为 82%,分别是 OECD iLibrary、IMF eLibrary、AIpatent、WorldWideScience、丝路科技知识服务系统、石油石化大数据知识服务平台、WDL、ICDL、Europeana。有 5 个平台支持按照“相关度”“发表时间”等维度对检索结果进行排序,占比为 45%,分别是 OECD iLibrary、IMF eLibrary、WorldWideScience、丝路科技知识服务系统、石油石化大数据知识服务平台。有 3 个平台支持二次检索,缩小检索范围,占比为 27%,分别是 IMF eLibrary、AIpatent、石油石化大数据知识服务平台。

所调查的平台中有 2 个可对检索结果进行可视化,占比为 18%,分别是 WorldWideScience 和石油石化大数据知识服务平台。WorldWideScience 可对检索结果的主题聚类结果进行可视化,深层揭示该检索结果下各主题的共现频次与分布规律。石油石化大数据知识服务平台能从发文量、关键词、学科、研究层次、文献来源、机构、作者、基金等维度对选定检索结果进行计量可视化分析。

4.2.5 界面与检索支持的语种

多语言界面包括多语言的导航栏、按钮、列表、弹窗等重要页面组件。各跨语言检索平台面向不同母语背景用户的需求,提供多语言界面,用户可直接在网站主页切换界面语种。在调查的平台中,有 7 个平台支持超过 1 种语种的界面,占比为 64%,分别是 OECD iLibrary、IMF eLibrary、AIpatent、丝路科技知识服务系统、石油石化大数据知识服务平台、WDL、ICDL 和 Europeana。其中 Europeana 的主要功能是向欧洲大众传播欧洲历史文化和科学知识,其界面语种版本覆盖欧洲大多数国家语种;AIpatent 整合的专利资源主要是日本、美国以及中国的官方专利数据库,其提供的界面语种版本是日语、英语和中文。

在跨语言检索方面,9 个平台支持不少于 3 种语种进行检索,占比为 82%,分别是 OECD iLibrary、IMF eLibrary、AIpatent、WorldWideScience、丝路科技知识服务系统、2lingual Google Search、WDL、ICDL 和 Europeana。各平台检索支持语种主要集中在常用语种,如中文、英语、俄语、法语、西班牙语、阿拉伯语、日语、葡萄牙语等。其中 2lingual Google Search 作为国际跨语言搜索引擎,不断增加其检索支持语种,2004 年发布平台原型时仅支持 11 种检索语种,目前已支持 37 种检

索语种。

以上典型的跨语言检索平台主要采用元数据层面的文献翻译方法和提问式翻译方法；跨语言翻译实现方法主要使用机器翻译方法，尤其是神经网络机器翻译技术实现跨语言翻译；平台提供简单检索和高级检索功能；能对检索结果进行排序和范围调整，并使用可视化技术呈现检索结果；其界面与检索支持常用语种，并不断扩展。

5 "一带一路"多语种共享型数据库的跨语言检索功能开发策略

"一带一路"多语种共享型数据库是一个多主体参与、多源异构资源归集、多语种覆盖的共享型数据库。以上调查结果能为"一带一路"多语种共享型数据库的跨语言检索功能设计与开发提供参考，具体如下：

5.1 采用基于神经网络机器翻译的提问式－文献翻译方法

在翻译方法上，"一带一路"多语种共享型数据库可借鉴现有跨语言检索平台的文献翻译方法和提问式翻译方法，采用两者结合的提问式－文献翻译方法。首先将源语言的提问式翻译成与待检索文献一致的源语言形式，进行单语言检索，然后将检索结果全部或部分翻译成由源语言描述的信息，该方法是目前实现跨语言检索比较理想的方法；在实现技术上，可借鉴Google、搜狗海外搜索和AIpatent，采用以神经网络机器翻译为主的机器翻译技术，其作为人工智能翻译主流技术[34]，能通过训练一张从一个序列映射到另一个序列的神经网络，输出变长的序列，相比于其他机器翻译技术，在翻译、对话和文字概括方面效率较高[35]；同时，神经网络机器翻译的开源工具丰富，为跨语言翻译系统构建和自动评价提供了平台基础和开发规范[36]。"一带一路"多语种共享型数据库使用神经网络机器翻译方法，可应用更先进的技术训练模型，优化神经网络结构，提高模型的表达能力，增加神经网络层数，进一步提升翻译质量和效率。

文献翻译方法可选择对结果文本的元数据、前两行、文摘或文本中重要的词语进行翻译。"一带一路"多语种共享型数据库资源类型丰富，尤其是手稿、历史资料、视频、图片、照片、地图、录音等仅具有条目的非文本馆藏，可借鉴中国台湾数字博物馆的跨语言信息检索实现策略[37]，通过对资源的元数据进行翻译，提供多语言元数据描述资源，有助于"一带一路"沿线国家不同母语背景的用户发现、识别、评价、选择和使用资源，实现资源的整合、共享、管理和长期保存。该方法充分利用了提问式翻译和文献翻译等优点，既简化翻译流程，降低用户的翻译成本，又提高了检索服务的质量。

5.2 实现多种检索功能

已建"一带一路"数据库多具备简单检索和高级检索功能，但极少提供专家检索功能。"一带一路"多语种共享型数据库应满足政府用户、企业用户、科研用户等不同专业水平用户的检索需求，提供简单检索、高级检索和专家检索功能，并制作数据库检索指南文件或在导航栏设立独立帮助中心栏目。在每个页面提供简单检索的一站式检索入口，用户可使用其统一检索入口检索数据库的异构资源；用户输入检索式后，需选择使用的源语言或平台能自动识别用户使用的源语言。在简单检索框旁应设有高级检索功能的链接，使用户能自由切换，设置高级检索功能为用户跨语言检索提供多种元数据的组合，提高跨语言检索查准率。设置专家检索，是因其功能强大，用户能通过构建布尔逻辑表达式进行检索提问，对检索结果有更准确的定向和控制作用，跨语言的专家检索方便用户使用自己熟悉的语言构造检索式，从而提高检索效率。

5.3 应用可视化技术呈现检索结果

可视化技术使检索过程更透明化，对检索结果进行形象生动的、有意义的分类组织，可建立有效的用户反馈机制和交互机制。"一带一路" 多语种共享型数据库需通过可视化手段，展现多国别、多类型、多语种资源的关联关系和发展逻辑，满足用户深层次、个性化的信息需求。

可对"一带一路"多语种共享型数据库检索结果进行可视化，通过统计、聚类、关联分析等手段分析处理检索结果数据集合，并将检索结果集合转换为二维或三维图形，采用直观的交互式、动态可视化方式揭示多语种信息资源，可加大用户对信息的认知度，加强系统的亲和度，有利于帮助用户快速理解外语资源，揭示信息资源的内在联系和深层含义。可借鉴周笑盈和魏大威梳理的演进描述可视化方法完善检索结果浏览功能，在时间线和地图上对"一带一路"信息资源实现时空维度的叙事可视化[38]。可参考孙倩、孙雨生、阮光册、邱均平等梳理的信息可视化关键技术完善检索结果分析功能[39－42]，提供合适的视图形式和层次结构，对全部或批量检索结果进行可视化分析，帮助用户快

速掌握检索结果的发文量、发文时间、作者、主题、期刊、语种、资源类型等维度的分布情况，深入揭示文献知识结构，方便用户选择浏览。此外，检索过程可视化可以应用动态可视化检索与过滤技术，帮助用户在与检索系统交互时，能以可视的方式执行并跟踪检索步骤，系统实时提供信息反馈，支持检索策略控制，可减少用户跨语言检索的记忆负担。

5.4 提供多语言界面和资源

提供多语言界面能使用户更好地适应多语言环境。考虑广泛适用性，"一带一路"多语种共享型数据库可先支持常用语种的界面版本，如中文、英语、俄语、法语、西班牙语、阿拉伯语，再选择性地扩展更小众的语种界面版本；或能根据用户 IP 地址识别用户所在地的常用语言，自动转换数据库网站的语言界面，以符合用户的使用习惯；将具有不同区域和国家特点的内容（例如数字、时间、货币等）以当地语言的格式显示，减少平台使用过程中可能存在的理解歧义，并保持文化中立性。"一带一路"多语种共享型数据库实现多语言界面应该维持一个源程序版本，易于修改、维护和升级，不同语言版本的网页之间在结构和业务逻辑上保持一致，即不同语言版本网页之间的差异都集中在 UI 层[43]，在加入新的语言版本时无需重新编译，可方便地扩展新语言。

对于已有的"一带一路"多语言资源，如政府文件、统计数据、调查报告、多个语言版本的书籍等，需将资源的所有语言版本收录完整，在检索结果详情页中予以提供，以便用户直接选择所需语言版本进行下载。对于翻译工作量小的网站说明、在线展览介绍和信息资源项，可利用机器翻译系统和人工辅助校对进行全文翻译，省去用户自己选择机器翻译系统进行翻译的步骤，综合考虑所需的成本和翻译准确率，可仅提供中文、英语等常用语言的资源版本，减少语言隔阂，推动"一带一路"数据库"走出去"。

如何实现跨语言检索是"一带一路"多语种共享型数据库和平台建设亟待解决的重要课题。笔者调查了已建"一带一路"数据库的检索功能设置，分析"一带一路"多语种共享型数据库检索功能需求，并对 11 个典型跨语言检索平台进行调研，归纳跨语言检索方法、跨语言翻译实现方法、检索功能设置、检索结果呈现、界面与检索支持语种 6 个方面的特点，借鉴典型跨语言检索平台的优秀建设经验，为"一带一路"多语种共享型数据库建设的跨语言检索功能设计与开发提出对策：应采用基于神经网络机器翻译的提问式－文献翻译方法，实现多种检索功能，应用可视化技术呈现检索结果，提供多语言界面和资源。本文为"一带一路"多语种共享型数据库建设提供了理论参考，以期为"一带一路"沿线国家信息资源建设和整合提供载体支撑和技术保障，为"一带一路"沿线国家的相关学术研究提供全面的多语言信息服务。

参考文献：

[1] 已同中国签订共建"一带一路"合作文件的国家一览[EB/OL]. [2020-09-22]. https://www.yidaiyilu.gov.cn/gbjg/gbgk/77073.htm.
[2] 苏新宁. 信息检索理论与技术[M]. 北京：科学技术文献出版社，2004.
[3] 赵生辉，胡莹. 数字图书馆跨语言信息服务等级框架研究[J]. 情报科学，2020，38(12)：63-69.
[4] 王昊. 跨语言信息检索实现方法与关键技术探讨[J]. 情报杂志，2005(7)：46-49.
[5] 李培，武丽辉. 网上信息的跨语言检索[J]. 情报资料工作，2004(2)：71-74.
[6] 郭宇锋，黄敏. 跨语言信息检索理论与应用研究[J]. 图书与情报，2006(2)：79-81，84.
[7] 张素芳. 国外跨语言信息检索中的翻译歧义性问题研究综述[J]. 图书馆学研究，2006(6)：72-75，78.
[8] 司莉，贾欢. 2004～2014 年我国多语言信息组织与检索研究进展与启示[J]. 情报学报，2015，34(6)：662-672.
[9] I SWARYA P，RADHA V. Adapting hybrid machine translation techniques for cross-language text retrieval system[J]. Journal of engineering science and technology，2017，12(3)：648-666.
[10] 许明武，赵春龙. 国内语料库翻译学研究的名与实[J]. 上海翻译，2018(4)：3-9，94.
[11] RAHIMI R，SHAKERY A，KING I. Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework[J]. Information processing & management，2016，52(2)：299-318.
[12] 黄海，蒋烈辉，何红旗，等. 基于 IDA 的反编译中间语言设计[J]. 计算机工程与设计，2009，30(20)：4734-4737.
[13] ONIFADE O F W，IBITOYE A O J，MITRA P. Embedded fuzzy bilingual dictionary model for cross-language information retrieval systems[J]. International journal of information technology，2018，10(4)：457-463.
[14] VILARES J，VILARES M，ALONSO M A，et al. On the feasibility of character n-grams pseudo-translation for cross-language information retrieval tasks[J]. Computer speech & language，2016，36：136-164.
[15] 郭华庚，赵英. 跨语言信息检索研究与应用[J]. 现代情报，2008(9)：142-145.
[16] 孙玥莹，何彦青，吴广印. 基于领域知识库的科技术语信息匹配模型研究[J]. 情报科学，2019，37(8)：16-21.
[17] 于施洋，杨道玲，王璟璇，等. "一带一路"数据资源归集体系建设[J]. 电子政务，2017(1)：8-14.
[18] 戴艳清，刘杨庆. "一带一路"研究与决策支撑平台资源组织

策略研究[J]. 图书馆学研究, 2020(16): 64－70, 80.

[19] 严丹, 李明炎. 高校"一带一路"研究的信息需求和资源支撑体系构建[J]. 图书馆建设, 2018(8): 56－63.

[20] 严丹, 马吟雪. "一带一路"专题数据库的建设现状及开发策略研究[J]. 图书馆学研究, 2017(12): 40－47.

[21] 梁昊光, 张耀军. "一带一路"语言战略规划与政策实践[J]. 人民论坛·学术前沿, 2018(10):98－105.

[22] 李月婷, 司莉. 基于语义的多语言信息组织模式研究[J]. 图书馆论坛, 2016, 36(2): 13－19.

[23] OECD iLibrary[EB/OL]. [2020－10－27]. https://www.oecd-ilibrary.org/.

[24] I MF eLibrary[EB/OL]. [2020－10－27]. https://www.elibrary.imf.org/.

[25] AIPatent[EB/OL]. [2020－10－27]. https://www.aipatent.com.

[26] WorldWideScience[EB/OL]. [2020－10－27]. https://worldwidescience.org.

[27] 丝路科技知识服务系统[EB/OL]. [2020－10－27]. http://silkroadst.ikcest.org.

[28] 石油石化大数据知识服务平台[EB/OL]. [2020－10－27]. http://oil.cnki.net.

[29] 2lingual Google Search[EB/OL]. [2020－10－27]. https://2lingual.com.

[30] 搜狗海外搜索[EB/OL]. [2020－10－27]. https://overseas.sogou.com.

[31] World Digital Library[EB/OL]. [2020－10－27]. https://www.wdl.org.

[32] International Children's Digital Library[EB/OL]. [2020－10－27]. http://en.childrenslibrary.org.

[33] Europeana[EB/OL]. [2020－10－27]. https://www.europeana.eu/portal/en.

[34] 林倩, 刘庆, 苏劲松, 等. 神经网络机器翻译研究热点与前沿趋势分析[J]. 中文信息学报, 2019, 33(11): 1－14.

[35] 张文, 冯洋, 刘群. 基于简单循环单元的深层神经网络机器翻译模型[J]. 中文信息学报, 2018, 32(10): 36－44.

[36] ZHANG B, XIONG D Y, XIE J S. Neural machine translation with GRU-Gated attention model[J]. IEEE transactions on neural networks and learning systems, 2020, 31(11):4688－4698.

[37] CHEN H H. Global digital library development in the new millennium[M]. 北京: 清华大学出版社, 2001.

[38] 周笑盈, 魏大威. 数字人文背景下基于需求的知识可视化方法研究——以国图公开课的视频内容可视化为例[J]. 图书馆, 2020(1): 20－28.

[39] 孙倩. 数字图书馆网站建设视角下资源可视化揭示的实践探索[J]. 图书馆理论与实践, 2017(5): 84－87.

[40] 孙雨生, 李万蓉. 国内数字图书馆信息可视化研究进展:架构体系与关键技术[J]. 图书馆学研究, 2019(4): 2－9.

[41] 阮光册, 任金玥. 基于主题层次关系的文献检索结果可视化应用研究[J]. 图书馆杂志, 2019, 38(5): 71－78.

[42] 邱均平, 余厚强, 吕红, 等. 国外馆藏资源可视化研究综述[J]. 情报资料工作, 2014(1): 12－19.

[43] 胡振宁, 杨巍, 丁培, 等. SULCMIS OPAC 多语言界面的设计与实现[J]. 现代图书情报技术, 2013(2): 70－76.

作者贡献说明：

司莉:确定文章整体思路及框架设计,修改论文;

周璟:数据库调研及数据采集,撰写论文初稿及修改。

Analysis and Development Strategy of Cross-Language Retrieval Function for "the Belt and Road" Multilingual Shared Database

Si Li　Zhou Jing

School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] To realize the effective use of "the Belt and Road" multilingual shared database resources, the problem of cross-language retrieval should be solved. Based on the survey results of "the Belt and Road" database retrieval function, "the Belt and Road" multilingual shared database's retrieval function demand is analyzed. From the perspective of researching on the cross-language retrieval platform, reference for cross-language retrieval function design and development of "the Belt and Road" multilingual shared database can be provided. [**Method/process**] Through literature and network survey, 11 typical cross-language retrieval platforms at home and abroad were selected. Analysis was carried out from five aspects: cross-language retrieval method, cross-language translation implementation method, retrieval function, retrieval results, interface and retrieval support language. Then concluded their implementation ways. [**Result/conclusion**] Based on this, strategies are proposed for the cross-language retrieval function design and development of "the Belt and Road" multilingual shared database: adopting question-document translation method based on neural machine translation, implementing multiple retrieval functions, visualization technology used to present retrieval results, providing multi-language interface and resources.

Keywords: "the Belt and Road" database　multilingual　cross-language retrieval